ЕКОЛОГІЯ ТА ОХОРОНА НАВКОЛИШНЬОГО СЕРЕДОВИЩА

УДК 556.531+519.24.001

Артеменко В.А., Петрович В.В., канд. техн. наук

ВІДНОВЛЕННЯ НАДВЕЛИКИХ ПРОПУСКІВ В ГІДРОЕКОЛОГІЧНИХ РЯДАХ МЕТОДОМ МУЛЬТИЛІНІЙНОЇ РЕГРЕСІЇ

Анотація. В статті наведений метод відновлення пропусків в часових рядах концентрацій речовин за даними повних рядів концентрацій інших речовин водного об’єкта.

Залежність між гідроекологічними параметрами описана з використанням лінійної функції багатьох змінних.

Запропонований метод дозволяє усунути безперервні пропуски в рядах даних довжиною декілька років.

Допустима істотна свобода у виборі опорних (повних) рядів та за величиною дискретизації їх за часом.

Відновлення надвеликих пропусків із використанням метода мультилінійної регресії показано на прикладі відновлення часового ряду розчиненого кисню у річковій воді та воді водосховища.

У даному випадку при відновленні пропусків для ряда розчиненого кисню у річковій воді було використано 14 предикторів, для водосховища – 13 предикторів.

Лінійний коефіцієнт кореляції між відновленими та дійсними значеннями завжди був більше 0.7.

Методика дозволяє також усувати надвеликі пропуски в різних природних часових рядах.

Ключові слова: гідроекологічні часові ряди, надвеликі пропуски, метод мультилінійної регресії, відновлення пропусків, часовий ряд розчиненого у воді кисню.

Abstract. A method has been developer to restore the time series of the concentration of the test substance from the results of observations of other time series of water body. Relationships between hydroecological components were determined using a linear function of many variables.

The proposed method can restore complete gaps in the source data, exceeding several years in time.

You can select the types of reference series and time slicing intervals.

The possibility of eliminating long gaps using multilinear regression is shown by the example of restoring the time series of dissolved oxygen in river water and water reservoir.

To restore the gaps in the time series, the concentration of dissolved oxygen in the river water was used 14 predictors, for the reservoir 13 predictors.

The linear correlation coefficient between the real and recovered values of the series always remained greater than 0.7.

In most cases, the considered method can adequately restore many natural time series.

Keywords: hydrological time series, very long caps in time series, multilinear regression method, restore the time series, time series of the concentration of dissolved oxygen.

Я буду цитировать гораздо
более достойную вещь – опыт,
наставника из наставников.
. . . Опыт никогда не ошибается.
Ошибаются только суждения ваши,
которые ждут от него вещей,
не находящихся в его власти.
Леонардо да Винчи

Введение

Как известно, процессы, происходящие в гидроэкологической системе, можно охарактеризовать при помощи различных временных рядов: гидрологических, гидрохимических, гидробиологических и др.

Практика гидроэкологического мониторинга обычно не обеспечивает достаточной частоты наблюдений. Наличие пропущенных значений в исходных

временных рядах значительно усложняет реализацию эффективных алгоритмов прогнозирования, поскольку такие алгоритмы в большинстве случаев требуют эквидистантных данных без пропусков [1]. Устранение пропусков в исходных данных и, соответственно, приведение их к эквидистантному виду представляют собой отдельные важные задачи [2,3].

Как правило, если пропуски небольшие (идущие подряд несколько точек), возможно использовать известные методы интерполяции: линейную, сплайнами и др. Однако если рассматривать устранение сверхбольших пропусков в несколько тысяч подряд идущих значений, такие пропуски обычными методами интерполяции устранить не удаётся. Прямым прогнозированием устранить их также не получается, так как большинство природных рядов прогнозируются всего на несколько точек вперёд.

В [4] был приведен метод восстановления пропусков в рядах одного водного объекта по полным рядам другого водного объекта, где в качестве исходных данных фигурировали только гидрологические ряды.

В развитие этого метода выполнено восстановление временных рядов концентраций различных химических веществ по ряду расходов воды в реке [5].

Проведенные нами численные эксперименты по восстановлению временных рядов концентраций одних химических веществ по временным рядам концентраций других химических веществ в реке (рядов без пропусков) дали в основном результаты вполне удовлетворительного качества.

В настоящее время при решении различных задач экологии практикуется системный подход, рассматривающий множество взаимодействующих компонентов системы в целом [6,7].

Анализируя свойства отдельных компонентов такой системы, следует отметить, что все они находятся в определённой взаимосвязи с другими компонентами. Поскольку в любой гидроэкологической системе целый ряд параметров взаимосвязан, вполне возможно восстановление одних временных рядов по качественно другим временным рядам.

Как показали предварительные численные эксперименты, взаимосвязи между гидроэкологическими компонентами водного объекта (реки, водохранилища) вполне адекватно могут быть описаны с помощью линейной функции многих переменных. При этом может использоваться достаточно много опорных рядов (рядов без пропусков).

Ранее линейная функция многих переменных была использована, например, при решении задачи о классах качества вод [2], при оценке численности популяции трески в зависимости от солёности и концентрации кислорода в придонном горизонте [8], а также других задачах.

Данное исследование проведено с целью устранения сверхбольших пропусков в исходных гидроэкологических временных рядах с помощью линейной функции многих переменных.

В статье рассматривается практический подход к устранению сверхбольших пропусков (пробелов) во временных гидроэкологических рядах сроком один год или даже более.

МЕТОД МУЛЬТИЛИНЕЙНОЙ РЕГРЕСИИ

В случае многомерной линейной регрессии величина **Y** зависит от нескольких независимых переменных [9,10].

В матричной форме уравнение линейной многомерной регрессии может быть представлено в виде

$$\mathbf{Y} = \mathbf{X}*\mathbf{K}, \qquad \textbf{(1)}$$

где **Y** – вектор зависимой переменной (отклик), **X** – регрессионная матрица, **K** – вектор коэффициентов линейной функции.

При этом число столбцов регрессионной матрицы будет на единицу больше числа предикторов, а первый столбец регрессионной матрицы есть, очевидно, единичный вектор - столбец.

Численные эксперименты, показали, что для нахождения коэффициентов можно использовать обычную регрессию (метод наименьших квадратов).

Коэффициенты уравнения **(1)** могут быть найдены по формуле

$$\mathbf{K}=(\mathbf{X}^T*\mathbf{X})^{-1}*\mathbf{X}^T*\mathbf{Y}, \qquad \textbf{(2)}$$

где T – оператор транспонирования, “-1” означает инвертирование матрицы.

МЕТОДИКА ИССЛЕДОВАНИЯ

В качестве примера использования мультилинейной регрессии в задачах гидроэкологии рассмотрим результаты восстановления пропусков во временных рядах концентрации растворённого кислорода в воде.

Для выполнения расчётов были использованы среднесуточные данные по р. Десна (створ с. Летки) и Киевскому водохранилищу за период с 1995 по 2010 год включительно (без пропусков). Для упрощения вычислений значения данных за 29 февраля высокосных годов из рассматриваемых рядов исключались.

При проведении численных экспериментов из исходных данных искусственно удалялась часть значений по концентрациям растворённого в воде кислорода. Изъятые данные были использованы в дальнейшем для оценки качества регрессии (см. ниже).

Для восстановления пропусков во временном ряду концентрации растворённого кислорода в речной воде было использовано 14, для водохранилища – 13 предикторов.

Перечень используемых предикторов для восстановления пропусков в ряду концентрации растворённого кислорода в речной воде приведен в Таблице 1, водохранилища – Таблице 2. Подобный подход позволяет установить – можно ли прогнозировать значения конкретного показателя, зная величины остальных базовых показателей.

Коэффициенты линейной многомерной регрессии между содержанием кислорода и выбранными предикторами были определены согласно **(2)** для 12 – летнего участка временного ряда (1995 … 2006 гг.). Для всех коэффициентов регрессии вычислялись доверительные интервалы.

Как известно, классически доверительные интервалы для коэффициентов регрессии определяются на базе формул, основанных на нормальном законе распределения Гаусса. Это требует дополнительных проверок на нормальность отклика, остатков, а также выполнения ряда других трудоёмких операций. Без таких проверок не рекомендуется выполнять как саму регрессию, так и находить доверительные интервалы.

Таблица 1 – Предикторы, используемые при восстановлении пропусков во временных рядах концентрации растворённого кислорода в речной воде

Номер п/п	Наименование предиктора	Номер п/п	Наименование предиктора
X01	Щёлочность воды, ммоль/$дм^3$	X08	Концентрация NO_2^-, мг/$дм^3$
X02	Концентрация Cl^- мг/$дм^3$	X09	Концентрация NO_3^-, мг/$дм^3$
X03	Концентрация CO_2, мг/$дм^3$	X10	Окисляемость кислородная
X04	Цветность воды, град.	X11	Водородный показатель pH
X05	Железо общее, мг/$дм^3$	X12	Расход воды, $м^3$/с
X06	Концентрация HCO_3^-, мг/$дм^3$	X13	Концентрация SO_4^{2-}, мг/$дм^3$
X07	Концентрация NH_4^+, мг/$дм^3$	X14	Температура воды, град. Цельсия

Таблица 2 – Предикторы, используемые при восстановлении пропусков во временных рядах концентрации растворённого кислорода в водохранилище

Номер п/п	Наименование предиктора	Номер п/п	Наименование предиктора
X01	Жёсткость воды, ммоль/$дм^3$	X08	Концентрация NO_2^-, мг/$дм^3$
X02	Щёлочность воды, ммоль/$дм^3$	X09	Концентрация NO_3^-, мг/$дм^3$
X03	Концентрация Cl^-, мг/$дм^3$	X10	Окисляемость кислородная
X04	Цветность воды, град.	X11	Водородный показатель pH
X05	Железо общее, мг/$дм^3$	X12	Температура воды, град. Цельсия
X06	Концентрация Mn^{2+}, мкг/$дм^3$	X13	Мутность воды, мг/$дм^3$
X07	Концентрация NH_4^+, мг/$дм^3$		

Вместе с тем в современных монографиях по регрессии все постулаты относительно связи регрессии и нормальности исходных данных снимаются.

Поскольку гидроэкологические данные в большинстве своём не есть нормально распределённые, в работе был использован качественно другой подход, основанный на использовании метода непараметрического бутстрепа. Непараметрический бутстреп позволяет легко определять доверительные

интервалы для различных параметров и статистик при любом законе распределения. В данном случае доверительные 95% - е интервалы для коэффициентов регрессии вычислялись по методике [11].

Для восстановления пропусков по содержанию растворённого кислорода согласно **(1)** была использована вторая (независимая) часть ряда за 2007 … 2010 годы, т.е. имитировался 4 – летний период непрерывных пропусков.

При этом временные ряды предикторов и растворённого кислорода рассматривались с различными интервалами квантования по времени (1 сутки, 5 суток, месяц и пр.).

На рисунках 1 … 4 представлены некоторые результаты восстановления ряда концентрации растворённого кислорода при различных интервалах квантования для реки и водохранилища соответственно.

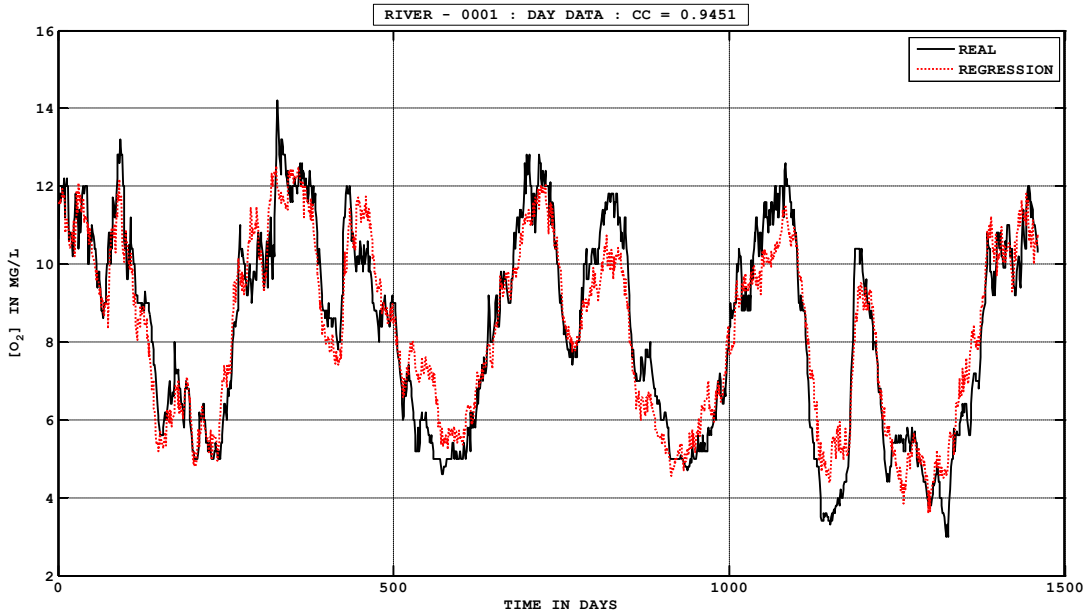

Рисунок 1– Результаты восстановления ряда концентрации растворённого кислорода в воде р. Десна за 4 – летний период (интервал квантования 1 день).

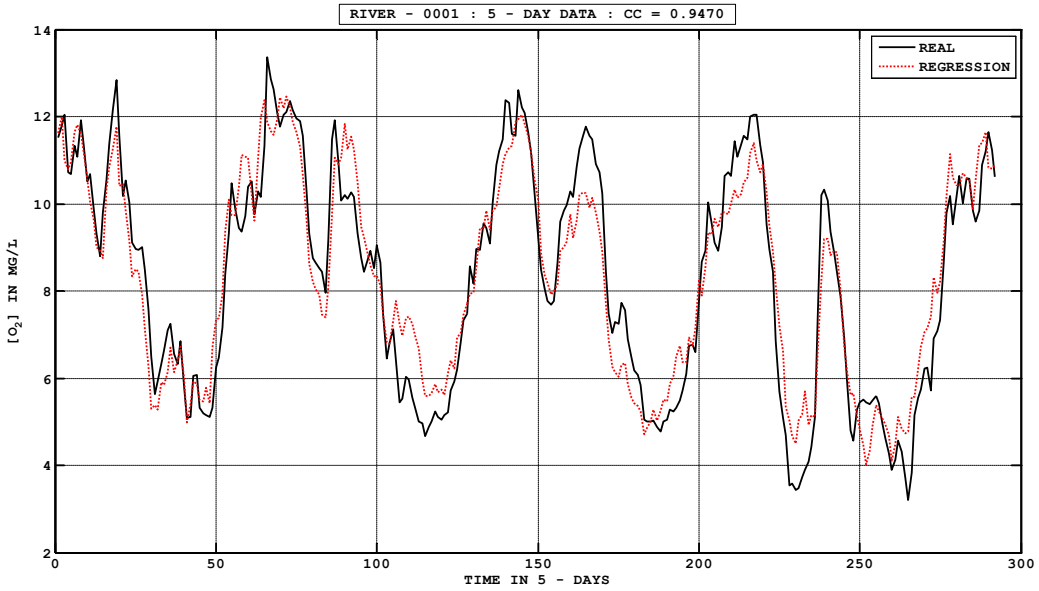

Рисунок 2 – Результаты восстановления ряда концентрации растворённого кислорода в воде р. Десна за 4 – летний период (интервал квантования 5 дней).

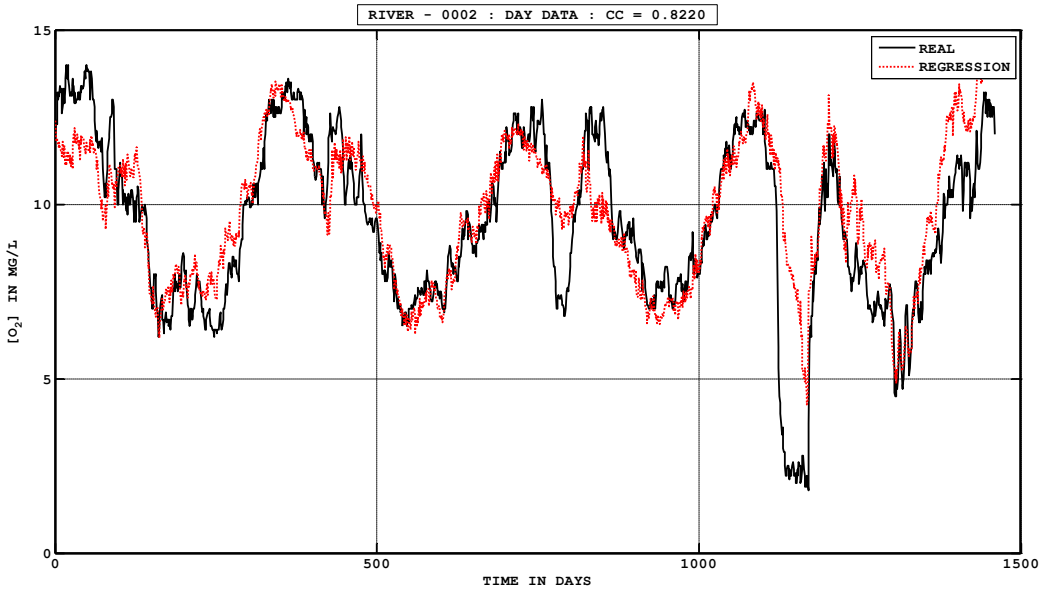

Рисунок 3 – Результаты восстановления ряда концентрации растворённого кислорода в воде водохранилища за 4 – летний период (интервал квантования 1 день).

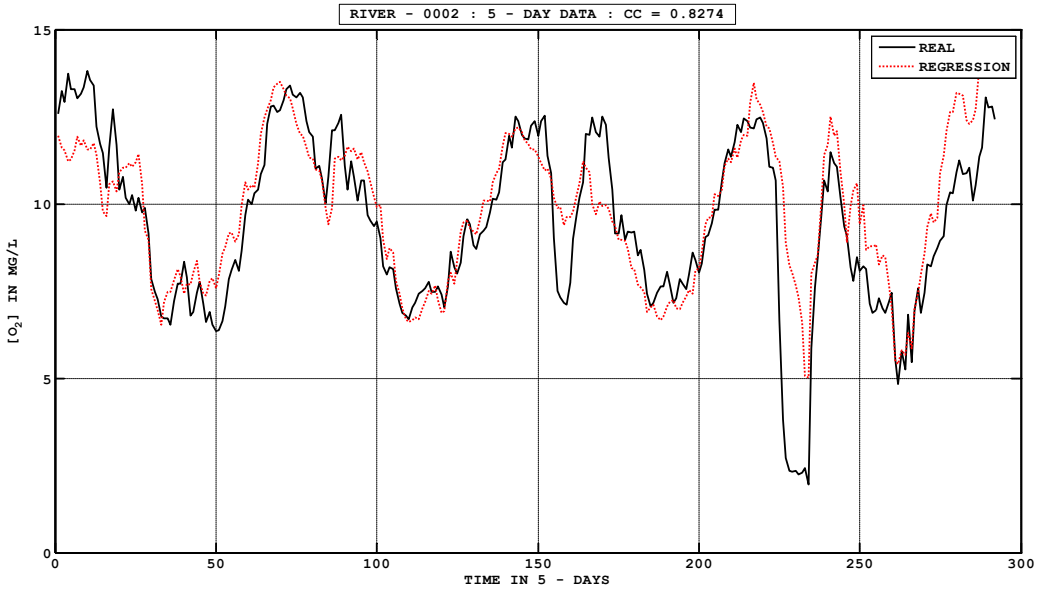

Рисунок 4 – Результаты восстановления ряда концентрации растворённого кислорода в воде водохранилища за 4 – летний период (интервал квантования 5 дней)

Следует особо отметить, что доверительные интервалы для некоторых коэффициентов регрессии содержат “внутри себя” значение “ноль”. Если строго следовать теории, полученные результаты регрессии будут не значимы. Но, поскольку оценивалось качество регрессии на независимом участке ряда, и качество оказалось достаточно высоким, то, несмотря на ряд ограничений теоретического плана, на практике вполне возможно использовать данный подход к построению многомерной линейной регрессии и получать положительные результаты.

Качество регрессии оценивалось с использованием классического линейного коэффициента корреляции Пирсона (CC).

Вопрос о том, что зависимость вида **(1)** действительно реальна, выясняется исключительно при восстановлении ряда растворённого кислорода по имеющимся опорным рядам. Принято считать, что такая зависимость существует, если ряд реальный, данные по которому у нас также есть (см. выше), и ряд восстановленный имеют коэффициент корреляции, равный 0.7 или более.

Как показали результаты расчётов, коэффициенты корреляции между реальными и восстановленными значениями остаются всегда существенно больше, чем значение 0.7.

Так, для реки величина СС, полученная на независимом участке ряда, составила 0.94 при интервале квантования 1 день, что значительно превышает критерий удовлетворительного результата.

В этой связи можно говорить о том, что исходные данные с точки зрения временного ряда растворённого кислорода обладают определённой структурой, которая хорошо “улавливается” при процедуре восстановления.

Также с помощью данного метода было успешно выполнено восстановление временного ряда гидрокарбонат – ионов и некоторых других веществ.

Однако рассмотренный метод восстановления работает не для всех рядов. Так, восстановление временного ряда концентрации аммонийных ионов с удовлетворительным качеством выполнить не удалось.

В заключение заметим, что использование каких – либо специальных методов регрессии не смогло существенным образом повысить качество восстановления данных. При этом большинство таких методов являются достаточно сложными для практической реализации.

Программно метод мультилинейной регрессии (и бутстрепа) был реализован на языке программирования MATLAB в бесплатных программах OCTAVE и FREEMAT.

При этом возможна работа не только в операционной системе WINDOWS, но также в LINUX.

Для нахождения вектора неизвестных коэффициентов **K** использовался специальный оператор языка программирования MATLAB – “обратный слеш” (обратная косая черта \)**.**

Выводы

1. В статье рассматривается практический подход к устранению сверхбольших пропусков (пробелов) во временных рядах сроком один год или даже более**.**

Разработан метод восстановления временного ряда концентрации исследуемого вещества при наличии результатов наблюдений за другими (опорными) рядами водного объекта.

При этом взаимосвязи между гидроэкологическими компонентами могут быть адекватно описаны с помощью линейной функции многих переменных.

Допускается достаточная свобода в выборе типов опорных рядов и интервалов квантования по времени**.**

2. Возможности устранения длительных пропусков с помощью мультилинейной регрессии показаны на примере восстановления временного ряда растворённого кислорода в речной воде и воде водохранилища.

Линейный коэффициент корреляции между реальными и восстановленными значениями ряда всегда оставался больше значения 0.7.

3. В большинстве случаев рассмотренным методом можно адекватно восстанавливать многие гидроэкологические ряды.

Полученные результаты показали, что существенно различающиеся по режиму функционирования водные объекты хорошо описываются указанными мультилинейными соотношениями.

Следует особо отметить, что временные ряды концентрации некоторых химических веществ данным методом восстановить не удавалось (например, ряды концентрации аммонийных ионов).

Перечень ссылок

1. Розенберг Г.С. Экологическое прогнозирование (Функциональные предикторы временных рядов) / Г.С. Розенберг, В.К. Шитиков, П.М. Брусиловский. - Тольятти: Изд-во Института экологии Волжского бассейна РАН, 1994. – 182 с.

2. Шитиков В.К. Количественная гидроэкология: методы системной идентификации / В.К. Шитиков, Г.С. Розенберг, Т.Д. Зинченко. - Тольятти: Изд-во Института экологии Волжского бассейна РАН, 2003. – 463 с.

3. Артёменко В.А. Устранение пропусков в гидрохимических временных рядах методом многообразий малой размерности / В.А. Артёменко, В.В. Петрович // Автомобільні дороги і дорожнє будівництво, вип. 102. - К.: Вид. Націон. трансп. ун-ту. - 2018. - С. 250 - 267.

4. Артёменко В.А. Метод восстановления сверхбольших пропусков в гидрологических временных рядах / В.А. Артёменко, В.В. Петрович // Автомобільні дороги і дорожнє будівництво, вип. 93. - К.: Вид. Націон. трансп. ун-ту. - 2015. - С. 150 - 156.

5. Артёменко В.А. Метод восстановления пропусков в гидрохимических временных рядах / В.А. Артёменко, В.В. Петрович // Автомобільні дороги і

дорожнє будівництво, вип. 101. - К.: Вид. Націон. трансп. ун-ту. - 2017. - С. 192 - 201.

6. Джефферс Дж. Введение в системный анализ: применение в экологии. - М.: Мир, 1981. – 252 с.

7. Страшкраба М. Пресноводные экосистемы. Математическое моделирование / М. Страшкраба, А. Гнаук. - М.: Мир, 1989. – 376с.

8. Дроздов В.В. Общая экология. - СПб.: Изд. РГГМУ, 2011. - 412 с.

9. Дрейпер Н. Прикладной регрессионный анализ / Н. Дрейпер, Г. Смит. - М.: Финансы и статистика. – 1986, Кн. 1, 366 с.; 1987. Кн. 2, 351 с.

10. Львовский Е.Н. Статистические методы построения эмпирических формул. - М.: Высшая школа, 1988. - 239 с.

11. Шитиков В.К. Рандомизация и бутстреп: статистический анализ в биологии и экологии с использованием R / В.К. Шитиков, Г.С. Розенберг. - Тольятти: Кассандра, 2013. - 314 с.